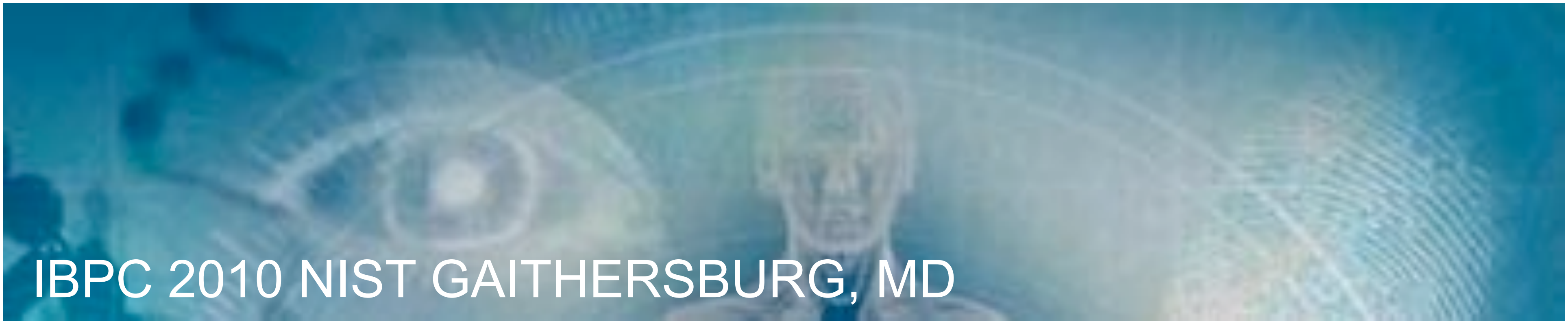


Biometric Covariate Analysis using Partial Area Under Curve

Valorie S. Valencia, PhD

President & CEO, Authenti-Corp

Research Professor of Optical Sciences
University of Arizona

March 4^{th} 2010

Collaborative Effort

Valorie Valencia — Authenti-Corp

Matthew Kupinski — College of Optical Sciences, The University of Arizona

Elham Tabassi — National Institute of Standards and Technology, U.S. Department of Commerce

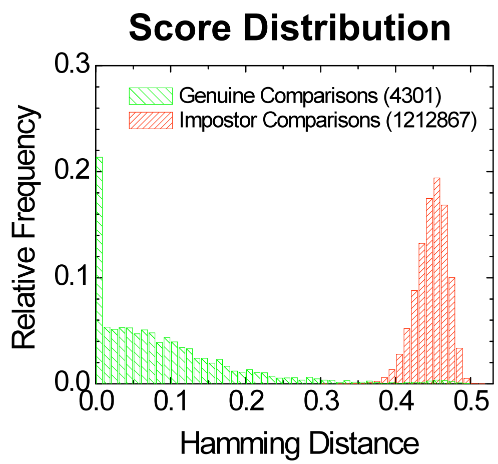
This work was performed under award 70NANB8H8145 from the National Institute of Standards and Technology (NIST), U.S. Department of Commerce. The statements, findings, conclusions and recommendations are those of the authors and do not necessarily reflect the views of NIST or the U.S. Department of Commerce.

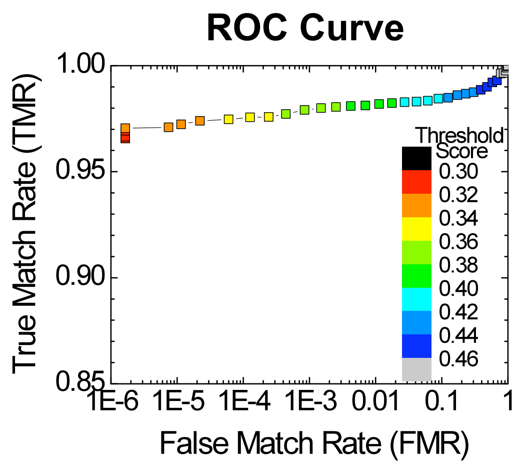
Why perform covariate analysis?

- It is important to understand the influence of various factors (covariates), such as image quality metrics, population demographic factors and environmental conditions on the performance of biometric recognition systems
- This knowledge can profoundly influence how biometric systems are designed and implemented in real-world operational scenarios
- To demonstrate our Area-Under-Curve method for performing covariate analyses, we explore matching performance for three iris datasets from Authenti-Corp's IRIS06 study using the Daugman 2007 algorithm

Covariate Analysis Challenges

- Biometric systems are used for many different types of applications, which necessarily operate at different points on an ROC curve.
  - For example, for admission to Disney World, the higher false match rates associated with lower false non-match rates (higher true match rates) would be tolerable
    - Convenience to the customer is more important than some level of monetary loss
  - At a high-security facility, the lower true match rates associated with lower false match rates would be required
    - Security is more important than convenience.
- The influence of covariates is typically analyzed at one or multiple operating points
  - For example, FMR=10^{-3}, 10^{-4}, 10^{-5} or Threshold Score=0.32, 0.34, 0.36 (Hamming distance)
  - Analysis at multiple points can be difficult, time consuming and cumbersome
  - Results can be difficult to convey and understand

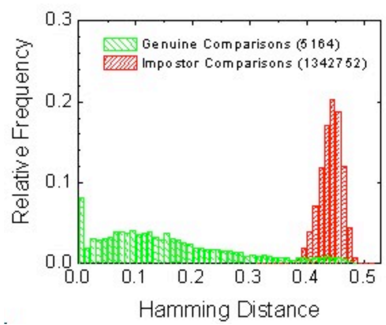




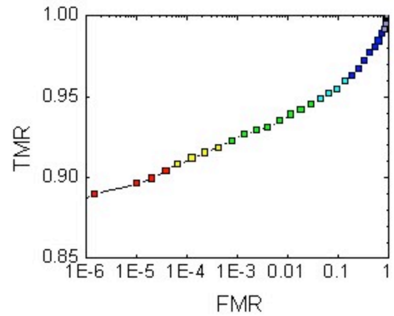
- It is desirable to perform a generalized covariate analysis that is independent of threshold ⇨ Area Under ROC Curve (AUC)

Why use Area Under Curve (AUC)?

Score Distribution



ROC Curve



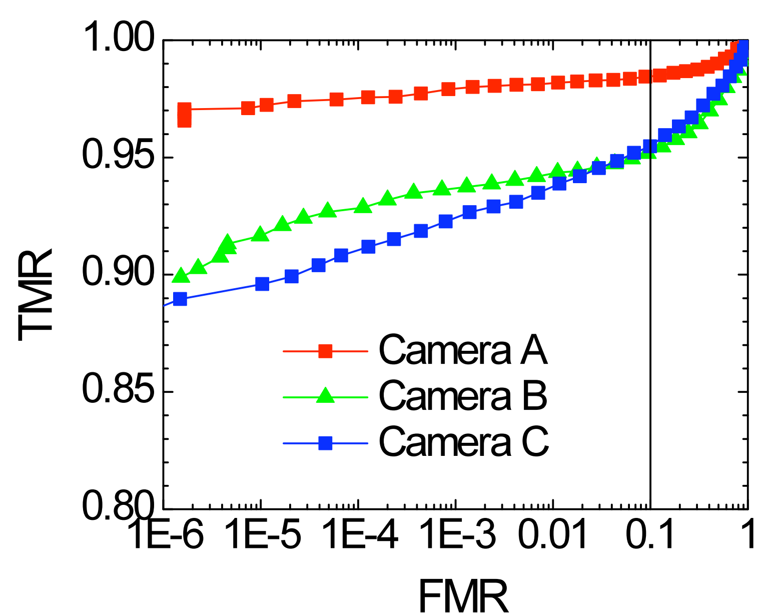
AUC

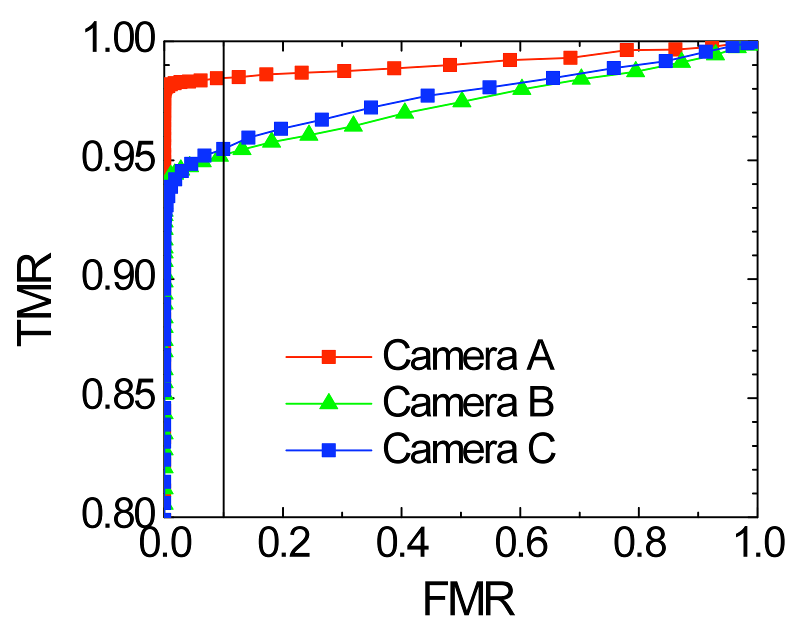
0.9763

- Easy to understand
  - Represents the probability of a correct decision given a genuine image and an impostor image
  - Overall probability of a correct answer
  - The larger the AUC value, the better the overall performance of the system
    - AUC=1 is perfect performance
- Serves as a single figure of merit that characterizes the performance of the system
  - Threshold independent
  - Accounts for all thresholds
- The statistical properties of AUC are well characterized
  - Determining statistical significance of AUC differences straightforward using Wilcoxon estimate
- The analysis space is reduced from a multi-point ROC curve to a single metric
  - The influence of various covariates on system performance can be systematically studied as a function of the AUC figure of merit

Limitations of AUC

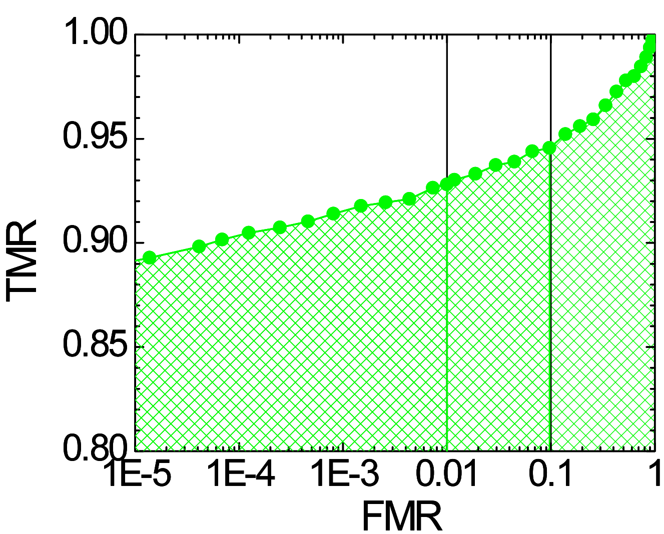
- Single metric from an inherently multi-objective problem
  - While problem is simplified, nuances may be overlooked
- AUC is heavily weighted by portions of the ROC curve where systems most certainly will not operate, that is at false match rates above a certain value, for example, FMR>0.1%

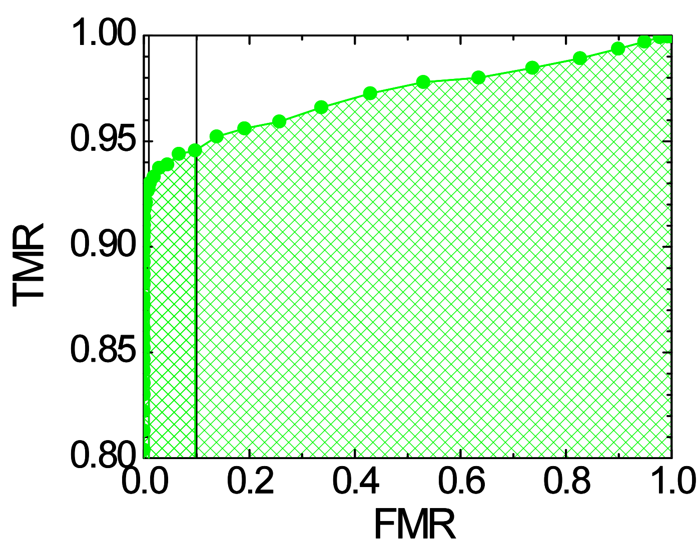




Partial AUC (p-AUC)

- To address limitations of AUC, we propose to look at partial AUC (p-AUC), which is restricted to a range of false match rates that are operationally feasible
- Selecting the range of the ROC curve that is operationally relevant depends upon the modality and scenario
  - For facial recognition, we have seen implementations that operate successfully at false match rates as high as 10%
  - For single-fingerprint systems, acceptable false match rates might be at or below 10^{-3}
  - For iris recognition, operational false match rates below 10^{-4} are typical





Area Under Curve
(probability of correct decision)

FMR ≤ 1.0, AUC=0.972292

FMR ≤ 0.1, p-AUC=0.093847

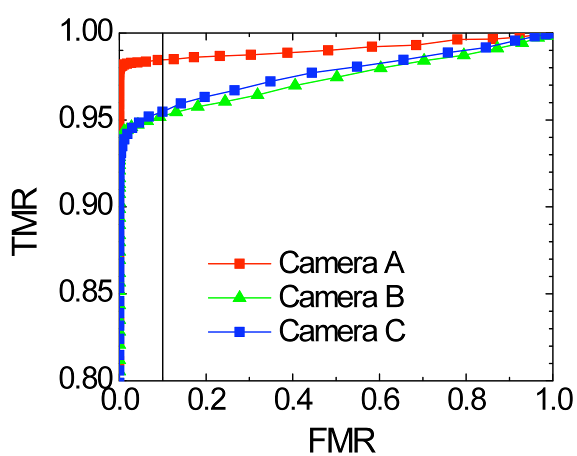
FMR≤0.01, p-AUC= 0.009219

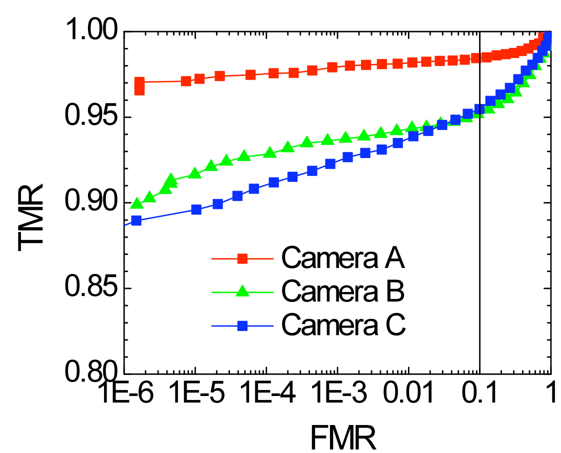
AUC Statistical Analysis

- Need error bars to draw conclusions
- Borrow image assessment approach from radiology
  - Probabilistic Multiple Reader, Multiple Case (MRMC) model
    - Normal cells ⇨ Genuine scores
    - Abnormal cells ⇨ Impostor scores
  - References
    - E. Clarkson, M. A. Kupinski, and H. H. Barrett, “A probabilistic model for the MRMC method. Part 1: Theoretical development”, *Acad. Radiol*., 13:1410-1421, 2006.
    - M. A. Kupinski, E. Clarkson, and H. H. Barrett, “A probabilistic model for the MRMC method. Part 2: Validation and applications”, *Acad. Radiol.*, 13:1422-1430, 2006.

Statistical Properties of AUC & p-AUC

- Variance $(\widehat{AUC})$ = $\sigma^2 = \dfrac{\alpha_1}{N_{gen}} + \dfrac{\alpha_2}{N_{imp}} + \dfrac{\alpha_3}{N_{gen}N_{imp}}$
  - Can directly compute each alpha term and predict variance from genuine and impostor scores
  - Third term accounts for correlations between impostors and genuines
  - OneShot freeware application computes α terms without resampling techniques and is unbiased http://www.radiology.arizona.edu/CGRI/IQ/page2/page7/page7.html
  - Methods and software extended to account for p-AUC





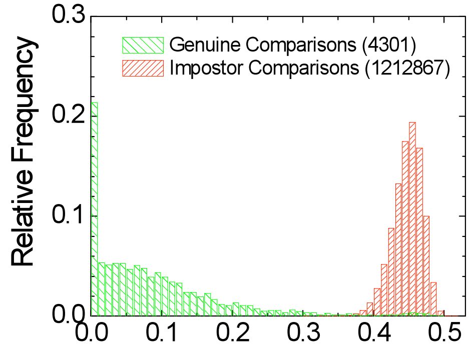
Statistical Significance "p-value"

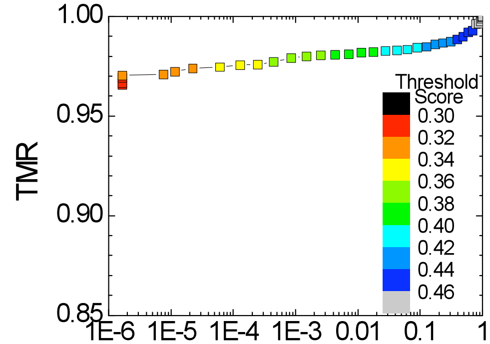
- Use Wilcoxon signed-rank statistical hypothesis test to determine statistical significance between two AUC values
- Non-parametric equivalent to t-test
- Assume null hypothesis ⇨ two AUCs equal
- p-value is the probability that the null hypothesis explains the result
  - Computed from the variances of the two AUCs
  - Small p-value (e.g., $p<0.05$) indicates a significant difference between the AUC values and thus a statistically significant performance difference between the two cases under investigation
- To perform the significance test for partial AUC, we assume that partial AUC is normally distributed
  - Normal assumption has been shown to be valid for as few as 10 subjects (i.e., 10 x 10 matrix of scores)
- Caution
  - p-value indicates statistical significance
  - p-value does not indicate that the hypothesis is correct

p-value Illustration

Score Distributions | ROC Curves | AUC ± σ

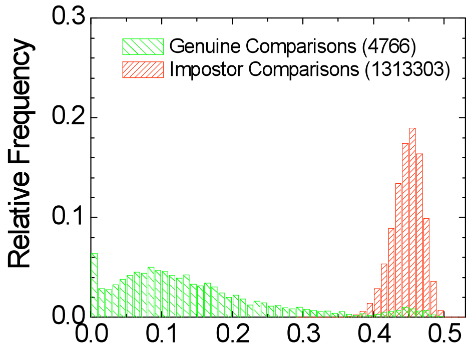
Camera A

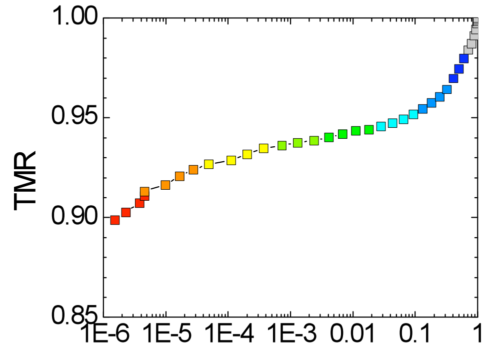




0.9905 ± 0.0012

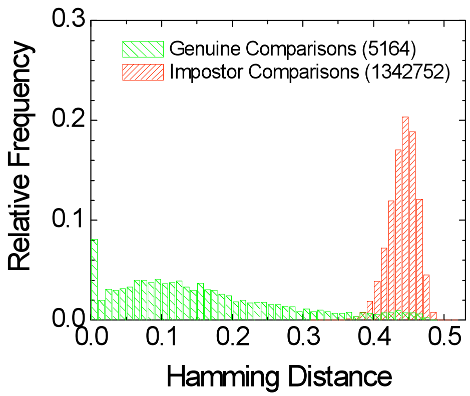
Camera B

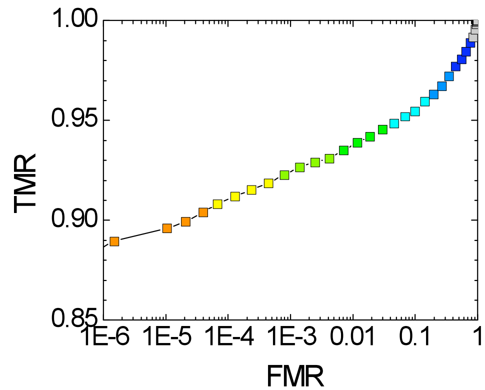




0.9730 ± 0.0018

Camera C





0.9763 ± 0.0016

Calculating p-value

Distribution of Measured AUC Difference (assuming true difference is 0) $= \dfrac{1}{\sqrt{2\pi\sigma_\Delta^2}} \exp\left[\dfrac{-1}{2\sigma_\Delta^2} x^2\right]$

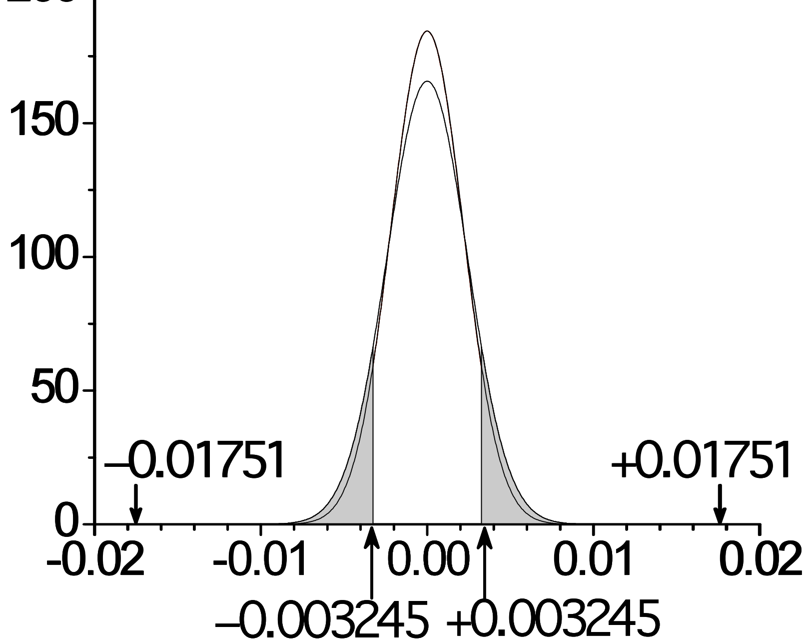
$$\Delta = |AUC_1 - AUC_2|$$

$$\sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

$\Delta_{\text{Camera B-Camera C}} = 0.003245$, $p = 0.1897$ Not statistically significant

$\Delta_{\text{Camera A-Camera B}} = 0.01751$, $p = 0.0000$ Statistically significant

Is the measured AUC difference unlikely?



Distribution of Measured AUC Difference
(assuming true difference is zero)

Integral form:

$$p = 2\int_{\Delta}^{\infty} \frac{1}{\sqrt{2\pi\sigma_\Delta^2}} \exp\left[\frac{-1}{2\sigma_\Delta^2} x^2\right] dx$$

Numerical form:

$$p = 2\left[\frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\Delta/\sqrt{2\sigma_\Delta^2}\right)\right]$$

probability of measuring observed difference if Δ=0

GLMM Covariate Analysis Approach

- Generalized Linear Mixed Effect model is used to relate probability of verification to subject and image covariates
  - Ross Beveridge's group at Colorado State University
- Pros:
  - Uses empirical performance and covariate data associated with people and imagery to fit a model relating covariate values to probability that a person will be correctly verified
  - Model quantifies how changes in covariates alter the probability that a person will be correctly verified
- Cons:
  - GLMM modeling complex
  - Requires parameter tuning
  - Performed at a selected operating point on the ROC, i.e., FMR=0.001

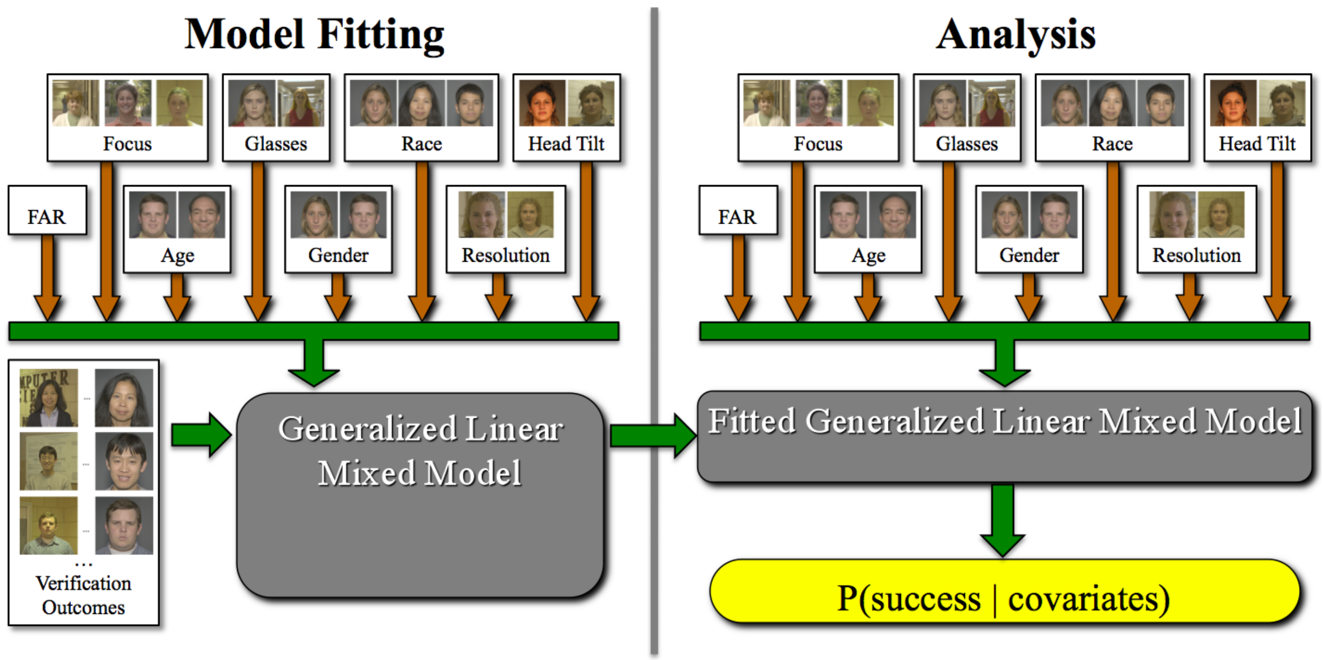


Figure from Beveridge, *et. al*., "Focus on Quality, Predicting FRVT 2006 Performance," 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition

AUC Covariate Analysis Approach

- To demonstrate utility of AUC & p-AUC figures of merit and Wilcoxon signed-rank statistical hypothesis test, we evaluate the influence of three covariates on iris recognition performance:
  – Camera
    - A, B & C
  – Gender
    - Male & Female
  – Eye
    - Left & Right

AUC & p-value Nomenclature

Camera

		A	B	C
	AUC	0.9905	0.9730	0.9763
A	0.9905		← p=0.0000	← p=0.0000
B	0.9730	↑ p=0.0000		↑ p=0.1897
C	0.9763	↑ p=0.0001	← p=0.1897	

probability of correct decision

- Camera A: 99%
- Camera B: 97%
- Camera C: 98%

p-value legend

← ↑ p>0.05, Not Statistically Signficiant

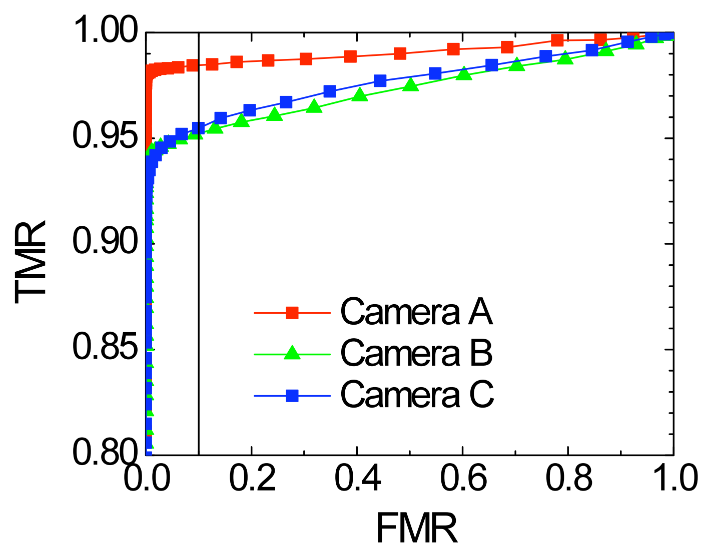
← ↑ $p \leq 0.05$, Statistically Significant

Direction of arrow indicates higher AUC value

Cameras A, B & C

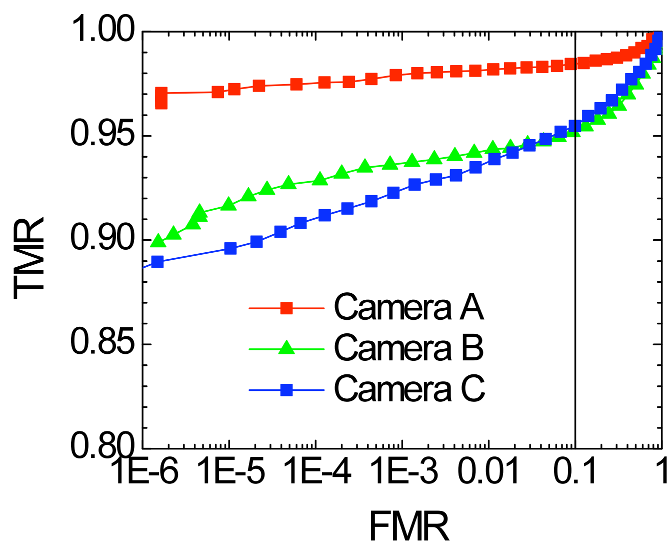
FMR ≤ 1.0

		A	B	C
	AUC	0.9905	0.9730	0.9763
A	0.9905		← p=0.0000	← p=0.0000
B	0.9730	↑ p=0.0000		↑ p=0.1897
C	0.9763	↑ p=0.0001	← p=0.1897	



FMR ≤ 0.1

		A	B	C
	p-AUC	0.0983	0.0948	0.0947
A	0.0983		← p=0.0000	← p=0.0000
B	0.0948	↑ p=0.0000		← p=0.7852
C	0.0947	↑ p=0.0001	↑ p=0.7852	



- Camera A performs significantly better than Cameras B & C
- Camera C performs better than Camera B for AUC (FMR≤1.0) but Camera B performs better than Camera C for p-AUC (FMR≤0.1)
  – ROC curves cross

Gender – Cameras A, B & C Combined

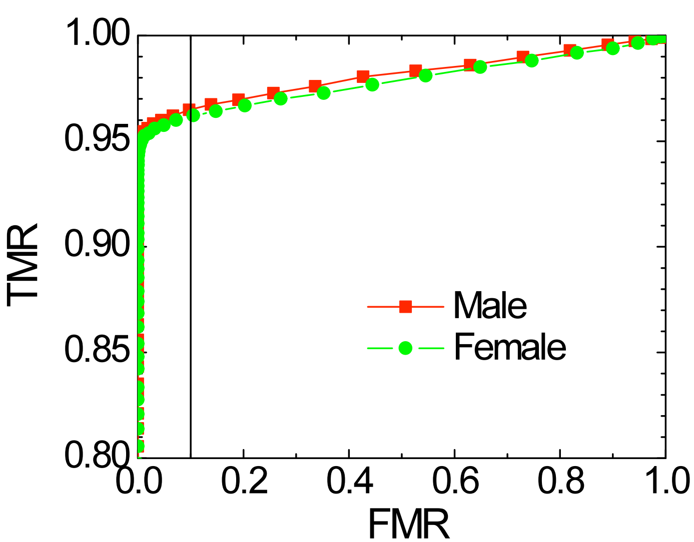
FMR ≤ 1.0

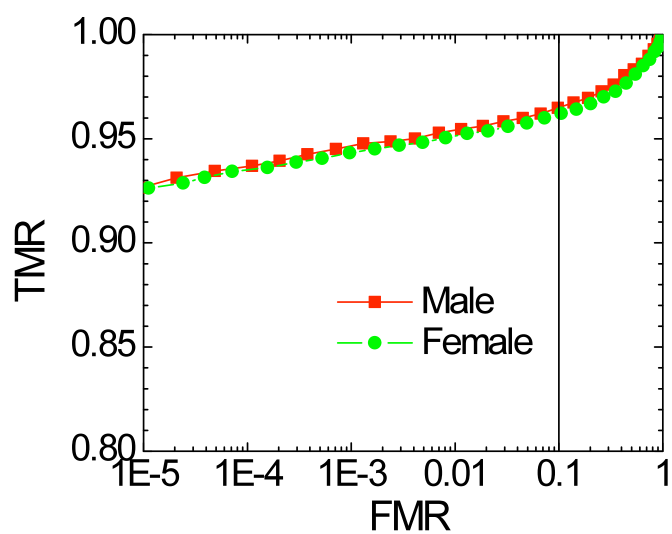
	AUC	Male	Female
	AUC	0.9814	0.9809
Male	0.9814		← p=0.1801
Female	0.9809	↑ p=0.1801	

FMR ≤ 0.1

	p-AUC	Male	Female
	p-AUC	0.0960	0.0957
Male	0.0960		← p=0.1897
Female	0.0957	↑ p=0.1897	

For Cameras A, B & C combined, there is no significant performance difference between men and women





Gender – Camera A

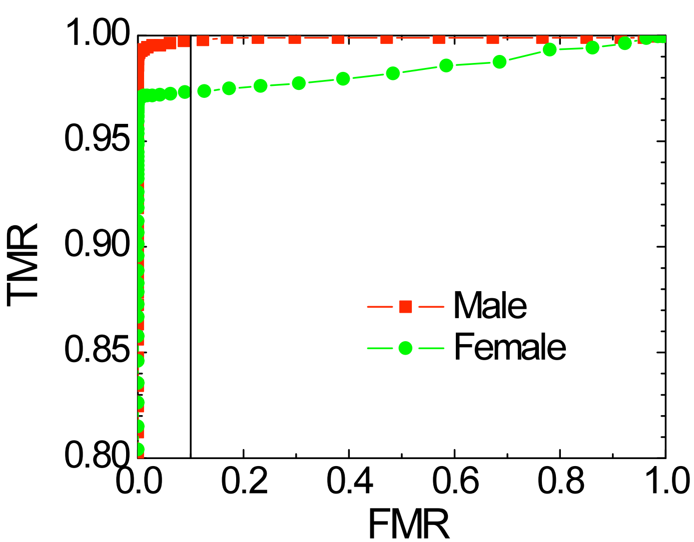
FMR ≤ 1.0

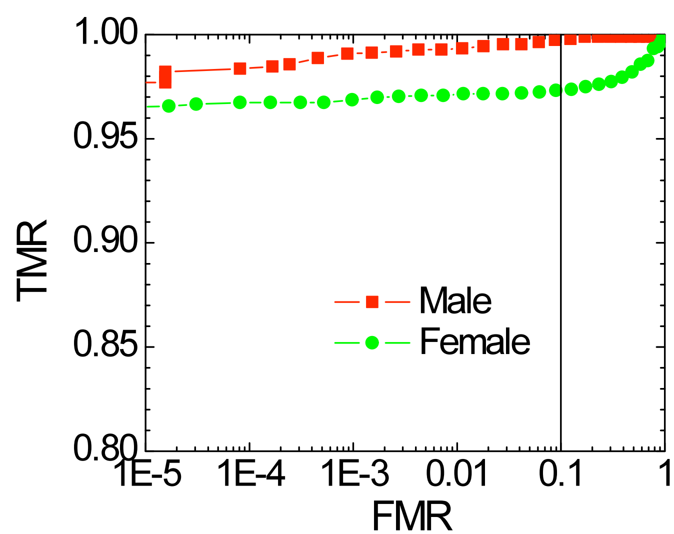
		Male	Female
	AUC	0.9986	0.9841
Male	0.9986		p=0.0000
Female	0.9841	p=0.0000	

FMR ≤ 0.1

		Male	Female
	p-AUC	0.0996	0.0972
Male	0.0996		p=0.0000
Female	0.0972	p=0.0000	

For Camera A, performance for men is significantly better than for women





Gender – Camera B

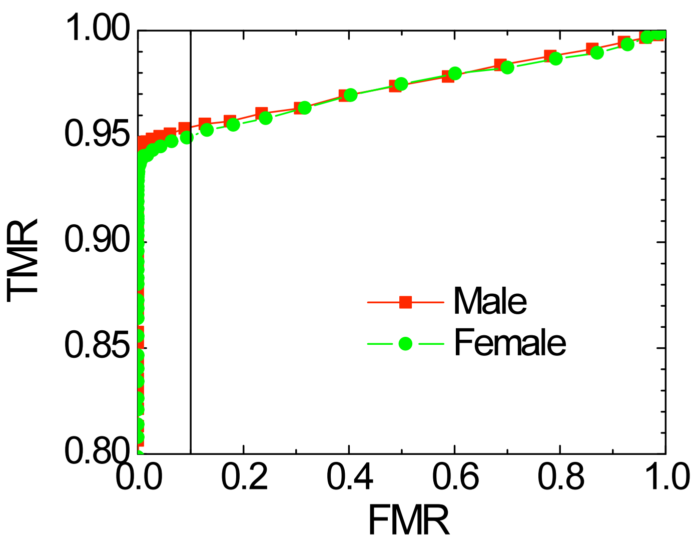
FMR ≤ 1.0

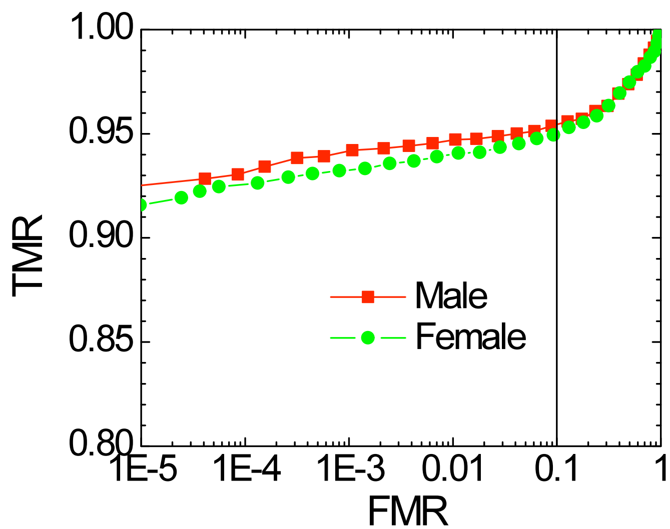
		Male	Female
	AUC	0.9742	0.9722
Male	0.9742		← p=0.5945
Female	0.9722	↑ p=0.5945	

FMR ≤ 0.1

		Male	Female
	p-AUC	0.0950	0.0945
Male	0.0950		← p=0.2421
Female	0.0945	↑ p=0.2421	

For Camera B, there is no significant performance difference between men and women





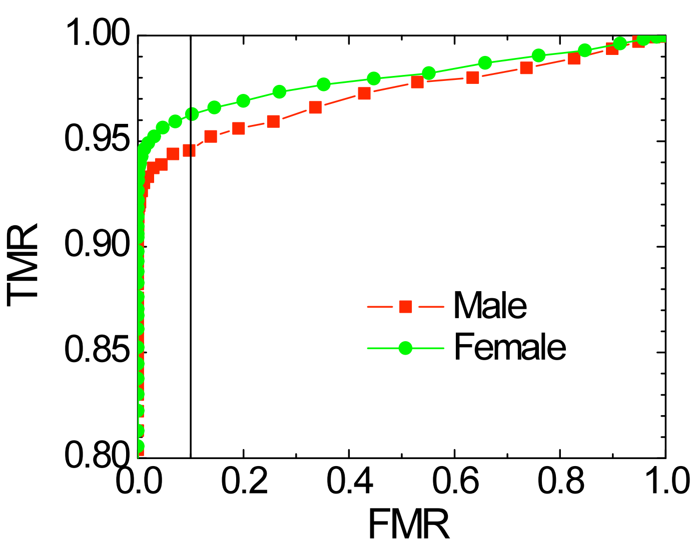
Gender – Camera C

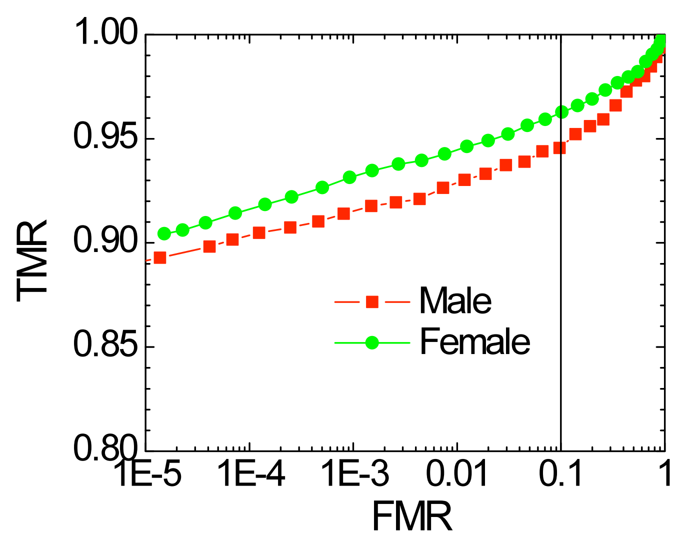
FMR ≤ 1.0

	Male	Female
AUC	0.9723	0.9797
Male 0.9723		↑ p=0.0276
Female 0.9797	← p=0.0276	

FMR ≤ 0.1

	Male	Female
p-AUC	0.0938	0.0954
Male 0.0938		↑ p=0.0001
Female 0.0954	← p=0.0001	





For Camera C, performance for women is
significantly better than for men

AUC and p-AUC figures of merit reveal performance variations between covariates

In this example:

- If population is predominantly male, use Camera A
- If population is predominantly female, use Camera C
- Can investigate origin of performance differences

Eye – Cameras A, B & C Combined

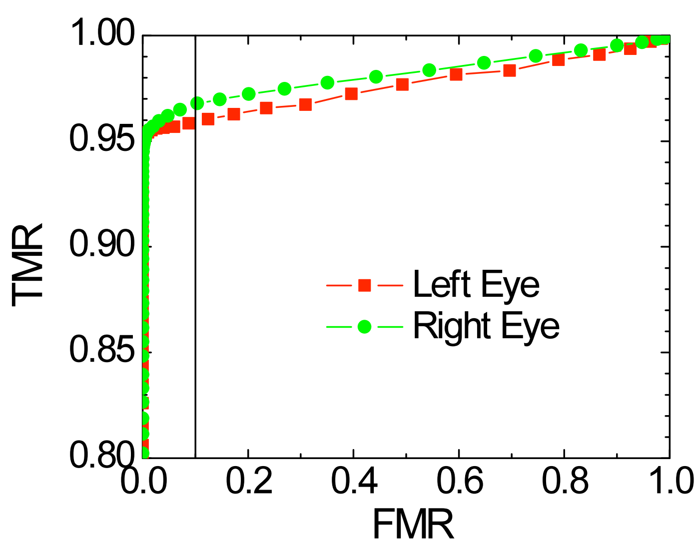
FMR ≤ 1.0

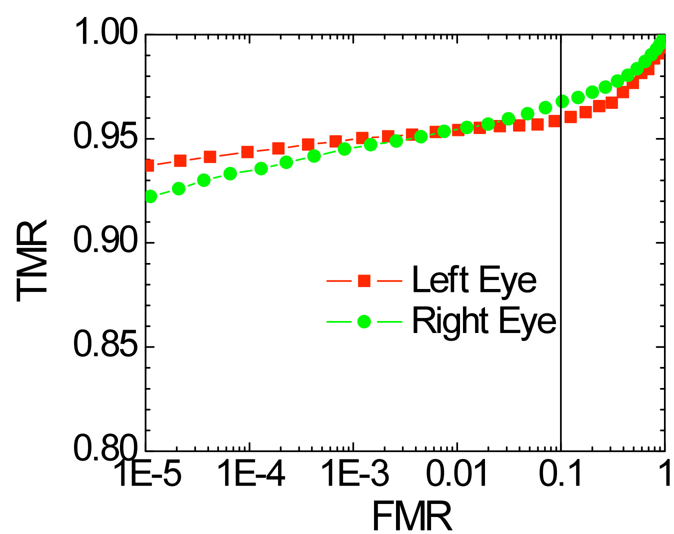
	AUC	Left	Right
	AUC	0.9776	0.9815
Left	0.9776		↑ p=0.0408
Right	0.9815	← p=0.0408	

FMR ≤ 0.1

	p-AUC	Left	Right
	p-AUC	0.0955	0.0961
Left	0.0955		↑ p=0.0059
Right	0.0961	← p=0.0059	

For Cameras A, B & C combined, performance for right eyes is significantly better than for left eyes





Eye – Camera A

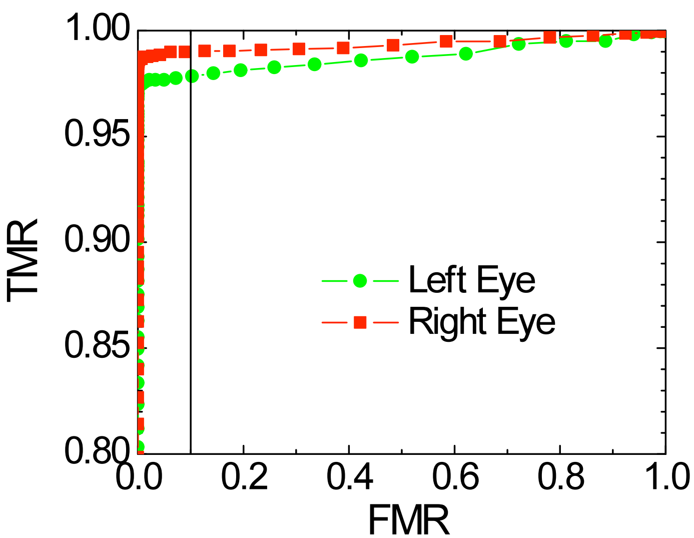
FMR ≤ 1.0

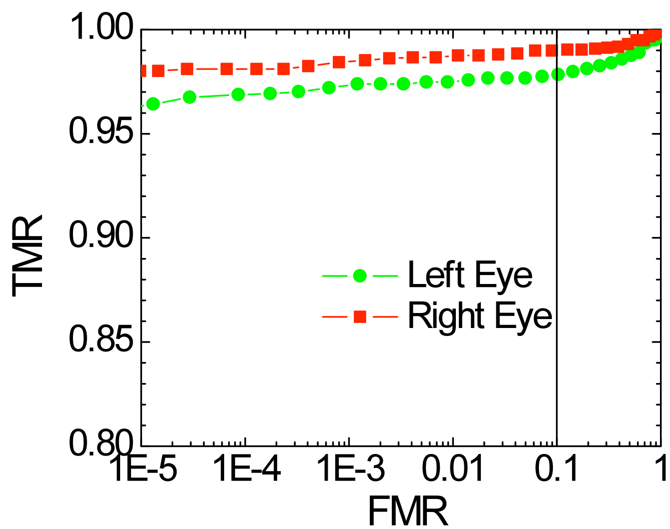
		Left	Right
	AUC	0.9875	0.9936
Left	0.9875		↑ p=0.0127
Right	0.9936	← p=0.0127	

FMR ≤ 0.1

		Left	Right
	p-AUC	0.0977	0.0989
Left	0.0977		↑ p=0.0000
Right	0.0989	← p=0.0000	

For Camera A, performance for right eyes is significantly better than for left eyes





Eye – Camera B

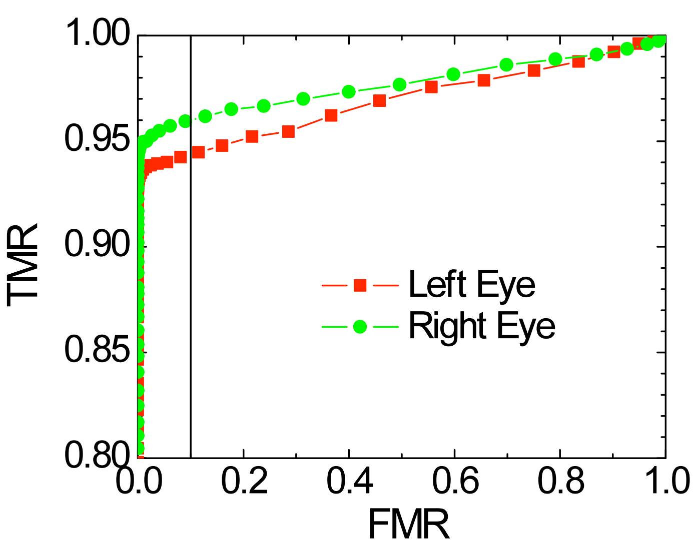
FMR ≤ 1.0

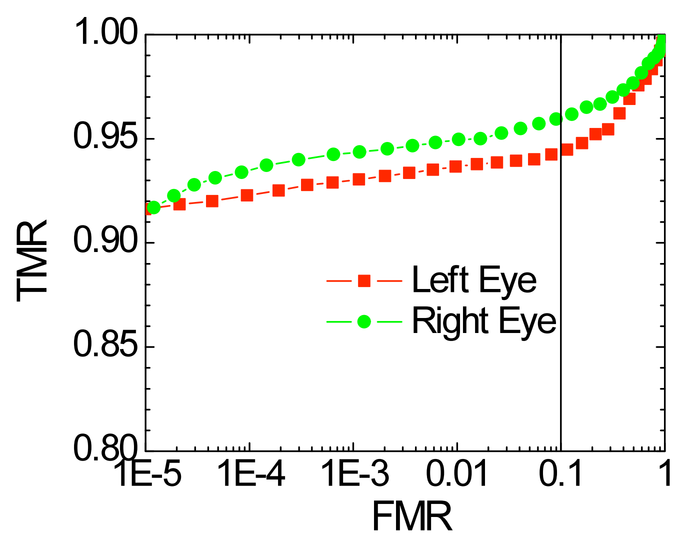
		Left	Right
	AUC	0.9692	0.9766
Left	0.9692		p=0.0439
Right	0.9766	p=0.0439	

FMR ≤ 0.1

		Left	Right
	p-AUC	0.0940	0.0955
Left	0.0940		p=0.0007
Right	0.0955	p=0.0007	

For Camera B, performance for right eyes is significantly better than for left eyes





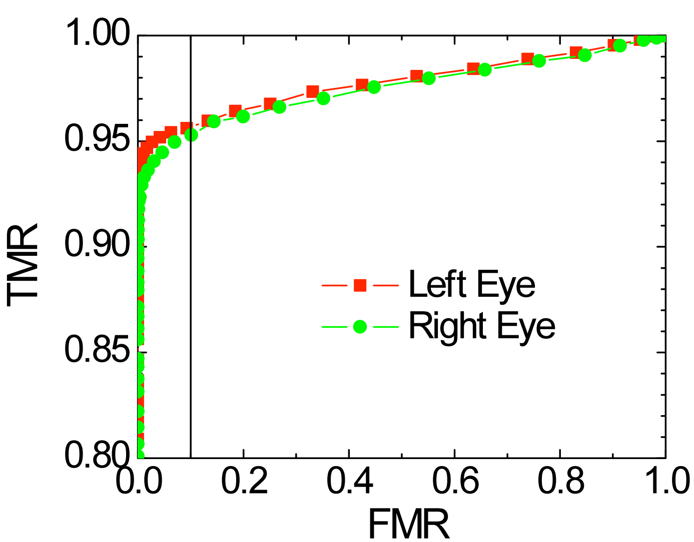
Eye – Camera C

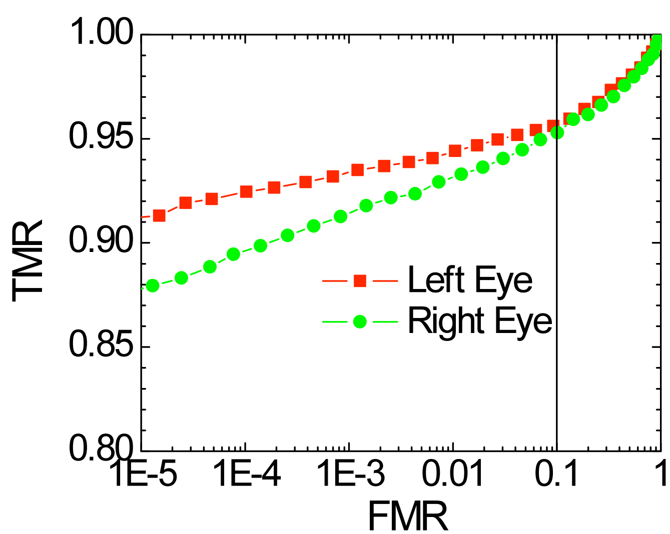
FMR ≤ 1.0

		Left	Right
	AUC	0.9778	0.9751
Left	0.9778		p=0.4259
Right	0.9751	p=0.4259	

FMR ≤ 0.1

		Left	Right
	p-AUC	0.0952	0.0942
Left	0.0952		p=0.0187
Right	0.0942	p=0.0187	





For Camera C, statistical significance is different for full AUC (FMR=1.0) and p-AUC (FMR=0.1)

- For full AUC there is no significant performance difference between left and right eyes
- For p-AUC, performance for left eyes is significantly better than for right eyes

p-AUC figure of merit reveals statistical significance for operational region of interest

In general, better to use p-AUC than AUC

Conclusions (1 of 2)

- Covariate analysis is an important tool for understanding the influence of various factors (covariates) and for enhancing the performance of biometric recognition systems
  - Identify which covariates matter and quantify how they affect performance for situations of interest
  - Useful to algorithm and hardware system developers
    - Facilitate system designs that are less sensitive or insensitive to significant covariates
  - Useful to system integrators
    - Implement systems to minimize influence of significant covariates
- Area Under Curve (AUC)-based covariate analysis approach is simple and fast to perform and easy to understand
  - AUC represents overall probability of a correct answer
  - Currently used in medical imaging field
  - System performance characterized with a single, threshold-independent metric
  - Re-sampling techniques not used
  - Produces unbiased estimates of components of variance
  - No modeling required, no parameters to tune

Conclusions (2 of 2)

- We propose a new metric, partial AUC (p-AUC), which is limited to an operationally-feasible portion of the ROC curve
- AUC and p-AUC are measures that give the probability of a correct decision when presented with both an impostor and a genuine image
- Statistical significance easy to determine using Wilcoxon p-values
  - Distribution of AUCs determines statistical significance of results
  - Small p-value indicates a significant difference between the metrics
- We have demonstrated the utility of AUC & p-AUC metrics and the Wilcoxon signed-rank statistical hypothesis test for performing covariate analyses using iris recognition data
  - The approach is effective, informative, straightforward and easy
  - Open-source code available for AUC
  - http://www.radiology.arizona.edu/CGRI/IQ/page2/page7/page7.html

Valorie S. Valencia, PhD

AUTHENTI-CORP

. . . providing scientific research, consulting, and evaluation services in all areas of authentication

PO Box 51675
Phoenix, Arizona 85076 USA
480-889-6444 office
602-432-0567 mobile
valorie@authenti-corp.com
www.authenti-corp.com

College of Optical Sciences
THE UNIVERSITY OF ARIZONA®
Frontiers of Biometrics Research

Meinel Building
1630 E. University Blvd
Tucson, AZ 85721-0094 USA
520-626-0155
valorie@optics.arizona.edu
www.optics.arizona.edu

Two-Reader Variance

$$\sigma^2 = \frac{\alpha_1}{N_{gen}} + \frac{\alpha_2}{N_{imp}} + \frac{\alpha_3}{N_{gen}N_{imp}} + \frac{\alpha_4}{2} + \frac{\alpha_5}{2N_{gen}} + \frac{\alpha_6}{2N_{imp}} + \frac{\alpha_7}{2N_{gen}N_{imp}}$$

$$\sigma_{12} = \frac{\alpha_5}{2N_{gen}} + \frac{\alpha_6}{2N_{imp}} + \frac{\alpha_7}{2N_{gen}N_{imp}}$$